■ 연구과제 요약문

과제명(기간)	안전한 인공지능 시스템 구현을 위한 적대적 공격 방어 기술 개발 (2019.03.01. ~ 2020.02.29.)
연구책임자	이 재 욱 (jaewook@snu.ac.kr)
개요	- 본 연구진은 본 연구과제를 통해 안전한 인공지능 시스템을 구축하고자 함. 최근 대용량 데이터 처리 기술과 딥러닝 기술의 발달로 인공지능 기반 정보처리 시스템이 증가하고 있음. 인공지능 시스템은 여러 이미지, 텍스트 등 다양한 고차원 비정형 데이터에 대해 뛰어난 성능을 보여줌. 그러나 데이터의 분포가 달라지거나 적대적 공격(adversarial attack)이 담긴 데이터에 취약하다는 단점이 있음. - 따라서 본 연구진은 비선형 사영법(non-linear projection), 임베딩(embedding) 기술, 도메인 적응(domain adaptation) 기술 개발, 그리고 강화학습 기반 순차적 커미티 머신(committee machine) 개발을 통해 어떠한 상황에서도 안정적인 성능을 보일 수 있는 견고한 인공지능 시스템을 개발하고자 함.
연구개발 결과	딥러닝을 활용한 클래스 의존적 표현학습 기술 개발 - 딥러닝을 활용하여 적대적 공격 방어 시스템 구축을 위한 표현학습 기술을 개발함. 적대적 공격을 방어하기 위해서는 같은 클래스의 샘플을 비슷한 위치로 임베딩을 시켜 적대적 샘플이 들어와도 결정경계를 넘지 않게 임베딩을 진행해야함. Sliced Wasserstein을 활용한 적대적 공격 방어 기술 개발 - 적대적 공격 방어 시스템 개발을 위해, 기존의 적대적 학습 방법에 새로운 제약식을 추가하여 적대적 샘플이 들어오더라도 안정적으로 분류를 할 수 있는 방법론을 개발함. 경사 다양성 정규화를 이용한 적대적 강건성 확보 - Randomized neural network을 공격하였을 때의 행동을 분석하고 이 분석을 통해 적대적 공격에 강건한 모델을 개발하고자 함.
활용분야 및 기대효과	- 인간의 눈으로는 구별 불가능한 작은 차이만을 주어 잘 학습된 모델을 높은 확률로 속일 수 있는 적대적 공격에 대한 효과적인 방어 전략을 구축하는 데에 큰 도움을 줄 것으로 예상됨. - 적대적 공격에 사용되는 적대적 샘플을 매니폴드에 사영시키면 매니폴드 상에서 데이터를 multi-basin 탐색이나 동적 시스템, 그리고 딥러닝을 통하여 효과적으로 방어할 수 있을 것으로 예상됨. - 1차년도의 비선형 사영법을 활용한 임베딩 기술 개발의 경우, 적대적 공격뿐만 아니라 Generative Model 등의 on-data 분야에서 핵심 연구 분야 중 하나임. 따라서, 이를 활용하여 이상치 탐지, 사전 경고 시스템 등의 분야에서 추가 연구가 가능할 것으로 예상됨.